中国 SKA 区域中心跨洲际高速数据传输进展及展望

郭绍光[1,2*]，安涛[1,2]，徐志骏[1,2]，劳保强[1,2]，陆扬[1]，吕唯佳[1,2]，伍筱聪[1,2]

1. 中国科学院上海天文台 SKA 区域中心联合实验室, 上海 200030

2. 鹏城实验室 SKA 区域中心联合实验室, 深圳 518066

* 联系人, E-mail: sgguo@shao.ac.cn

收稿日期：2022-06-28；接受日期：2022-0x-xx；

国家重点研发计划 (编号:2018YFA0404603)、SKA 专项 (编号:2020SKA0110300)、国家自然科学基金 (批准号:11873079,12041301) 和中国科学院青年创新促进会项目 (编号:2021258) 资助项目

摘要　平方公里阵列望远镜（SKA）作为最大的射电望远镜，其观测产生的数据将首先由澳大利亚和南非两个台址国传输到百公里左右的科学数据处理中心，然后通过高速网络分发到上万公里距离的各个 SKA 区域中心。具有 SKA 10% 规模的 SKA1 阶段，每年预计有 750PB 的数据需要通过至少 100Gbps 的网络分发到各个 SKA 区域中心（SRC），如此高的网络带宽和数据规模对数据的传输分发带来极大挑战。本文通过对 TCP/UDP/HTTP 等不同网络协议的分析，并使用当前射电天文领域不同的软件进行测试和研究，得出了目前在 10Gbps 网络的基础设施下最佳的传输方案参数，文中讨论了影响高速传输的因素，给出了相应的性能优化的策略，在 SKA1 真正的观测数据产生之前，将为中国 SKA 区域中心的网络建设和布局提供技术基础。描述的技术细节和方法可供相关科学应用参考和使用。最后讨论并展望了未来 SKA 网络需求的挑战。

关键词　平方公里阵列, SKA 区域中心, 高速网络, 数据传输

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

1　引言

平方公里阵列望远镜（Square Kilometre Array, 简称 SKA）是一项国际大科学工程，是由全球多国合资建造和运行的世界最大规模综合孔径射电望远镜，也是由多个关键科学问题驱动的大科学工程，具有超大视场、超高分辨率和超高灵敏度的特点，其灵敏度预计提高 50 倍，巡天速度预计提高 10000 倍 [1–3]，SKA 的数据生成、传输、处理必然导致极高的数据流产生，也对从台站产生数据、传输到中央信号处理器 (Central Signal Processor，简称 CSP)、及最后到 SKA 区域中心（SKA Regional Centre，简称 SRC）等各个阶段的数据传输提出了很高的需求 [1,4–6]。根据目前的规划，SKA 第一阶段（SKA Phase 1，简称 SKA1）已于 2021 年 7 月正式启动建设，预计 2029 年底建成并投入运行，产生的数据需要同步传输到位于全球的各个 SRC，根据每个 SRC 数据分担的评估，为了满足 SKA1 的数据传输，网络带宽至少需要 100Gbps，才能保证 SKA1 产生的数据稳定及时地到达各个区域中

心，SKA1 正式运行后的 SRC 的数据分发量预计为每年 710PB，算力至少需要 22PFlops [4]。中国科学院上海天文台于 2019 年主持建设部署了第一个 SKA 区域中心原型系统（CSRC-P）[7]，用于开展 SRC 前期的科学预研究，并为后期 SRC 的建设提供经验基础。作为 SKA 科学数据处理系统的一个重要组成部分，考虑到其重要性，SKA 天文台（SKA Observatory，简称 SKAO）为此也专门成立了 SKA 区域中心指导委员会（SRC Steering Committee，简称 SRCSC），用于考虑 SRC 的相关需求。SRCSC 将 SRC 的网络建设分为两个阶段 [4] 来推进，第一个阶段为 2020 年至 2025 年，网络功能达到 80% 的能力，具备 SKA1 基线能力的 10%；第二个阶段为 2025 年至 2030 年，网络功能达到 100% 的能力，具备 SKA1 基线能力的 100%。这里描述的 SKA 区域中心基础设施建构类似于欧洲核子研究组织（Conseil Européenn pour la Recherche Nucléaire，简称 CERN）[8] 的全球大型强子对撞机计算网格（World Large Hadron Collider Compute Griding，简称 WLCG）技术，WLCG 采用的分层数据处理模型利用协议来确定区域中心需要执行的数据处理。SRC 网络将提供给 SKA 社区访问数据的权限及用于分析处理数据的工具和软件 [5]。本文基于CSRC-P 系统，讨论中国 SRC 需要具备的网络基础和相关政策，通过与目前 SKA 的先导项目默奇森宽视场阵列 (Murchison Widefield Array，简称 MWA)、澳大利亚 SKA 先导者 (Australian Square Kilometre Array Pathfinder，简称 ASKAP) 及南非 SKA 探路者项目 MeerKAT 等的合作来阐述CSRC-P数据传输的概况与将来的升级计划等。以此来探讨 SKA1 及全规模 SKA 正式运行后，中国 SKA 区域中心的跨洲际高速网络的应对措施和解决方案。

介绍了一种支持 SRC 全球分布站点需求的网络结构和设计，通过当前与不同的国家和机构开展的实测数据传输，来得出目前使用的协议和软件需要进行的参数优化，并对未来 SKA 区域中心数据传输提出了一些建议和规划。

2 SKA 网络和数据流

SKA 具有超大视场、超高分辨率、超高灵敏度的特点，同时也面临在新技术背后产生的海量数据的挑战，针对不同的科学目标，从天线数据的获取、存储，然后将数据在中央信号处理器（Central Signal Processor，简称 CSP）上进行在线或离线的相关处理，进一步到达科学数据处理器 (Science Data Processor，简称 SDP) 对数据质量进行检查，包括初始校准、快速成像等预处理，生成不同类型的科学产品，最后将预处理的数据传输到分布于各大洲的 SRC [9] 进行软件环境的搭建 [10]，科学数据的优化处理 [11–13]。

其中位于澳大利亚和南非的两个观测台站会将所有经过 SKA 科学数据处理器预处理后的科学数据发送到世界各地的 SRC，每个 SRC 需要保存数据副本的一部分 [14,15]，同时发送该部分到另一个 SRC 产生一个副本。该网络需要支持 SKA 数据从观测台站到 SRC 进行数据的移动，以及 SRC 之间的数据交换。

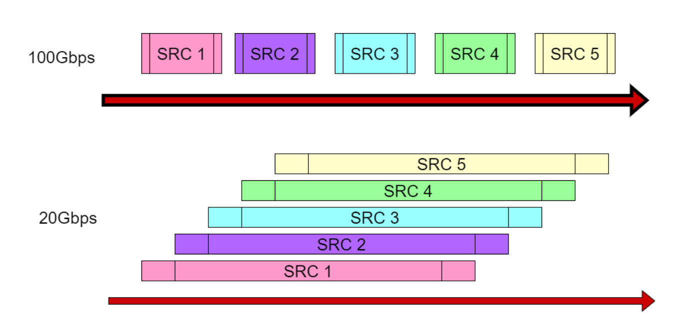


图 1 SKA1 数据交付模型图 [14]
Figure 1 Data Deliver Model of SKA1

海量的数据对网络、计算和存储都提出了新的需求，数据的传输、存储、归档和处理都面临很大的挑战。对于 SKA 科学数据处理系统（Science Data Processing，简称 SDP）而言，数据的处理可以在距离台站较近的地方进行，但此时的数据仅仅经过了最初的校准处理和质量检查，而后续的编辑、校准、成像、识别、工具开发、归档管理、基准测试、数据分析及相关科学均需要在 SRC 进行，所以，SKA 天文台的 SDP 或者 CSP 产生的数据并非最终的

科学分析数据，而且，如此大的数据流交付给最终的用户是不可行的，且 SKAO 也规定 SDP 只能由天文台的操作人员访问 [16]，科学用户只能从 SRC 获取 SKA 的科学数据 [4]，另外由于观测台站、区域中心以及最终科学用户分布在全球，需要涉及到 SKA 的信号传输与网络子部件（Signal Transport and Networks for the SKA，简称 SaDT），这些都对 SRC 的全球网络协作提出了很高的要求。一般来说，传输到 SRC 的数据产品将会是非常大的数据文件，单个文件在 100GB 左右，通过网络数据包往返时间（Round-Trip Time，简称 RTT）高达 100 ~ 300 毫秒的长距离路径传输到 SRC，这些连续的数据流需要稳定、无损的端到端网络，并且没有瓶颈限制。这里有两种网络的操作模型如图1所示，一种为通过 100Gbps 的网络带宽将 SKAO 的数据产品依次发送到每个 SRC，另外一种为数据产品同时向多个 SRC 并行传输，并发向 5 个 SRC 传输，每个数据链路仅需要 20Gbps 的网络带宽即可。

SKA 的详细的数据网络拓扑图如图2所示。台站产生的原始观测数据，在 SDP 经过初步处理，通过数据路由分发到每个 SRC 的子网络，天文学家通过互联网连接到区域数据中心进行深度数据处理，包括数据的分析、可视化、建模以及软件工具的开发，此时的数据称为高级数据产品（Advanced Data Products，简称 ADP），此时的数据将通过云的方式提供给科研人员。

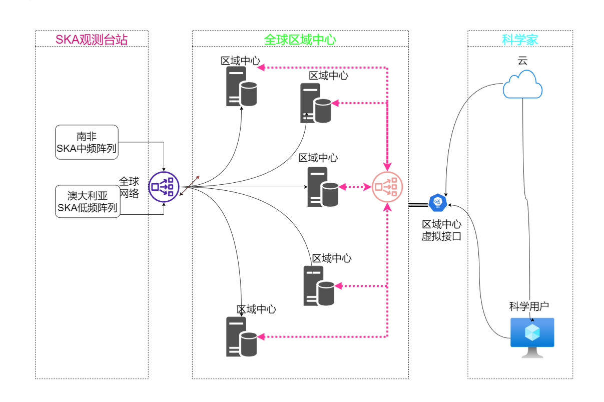


图 2　数据交付拓扑图
Figure 2　Data Deliver Tier

根据当前 SKA1 的规模可知，中频阵列单个天线的数据输出速率为 64Gbps；低频阵列单个台站的数据输出速率为 16Tbps，需要传输到区域中心的数据量在 710PB 左右 [4]。如图2所示，目前预估从 SKA1 产生并需要传输到每个 SRC 的数据流在 200 ~300PB/year，其最终处理后的数据通常会比这个数据量小，但是在未确定最终处理模型的情况下，多个模型对比产生的数据很可能会比原始数据还大。

以 2013 ~ 2015 年银河系和银河系外 MWA 巡天 (The GaLactic and Extra-galactic All-sky MWA SURVEY，简称 GLEAM) [17] 为例，该巡天的瞬时频率带宽为 30.72MHz，在 72MHz 至 231MHz 之间分为五个频段，为了避免太阳的影响，观测只在夜间进行，观测分为一系列的 120 秒的快照，故 10 分钟以内可以完成 5 个频段的观测，每次 2 分钟的观测时间所产生的原始可见度数据为 46GB，对其做 2 秒积分，平均 40kHz 频率分辨率，产生的数据在 26GB，最终的 coarse 通道的数据在 200MB 左右，如果将积分时间降低，频率分辨率提升，会直接增加产生的数据量。后续的 GLEAM 扩展巡天（THE GALACTIC AND EXTRA-GALACTIC ALL-SKY MWA EXTENDED SURVEY，简称 GLEAM-X）[18]，将阵列的天线数量增加了一倍，阵列的最大基线也增加了一倍，有效提高了 uv 覆盖，并将成像的信噪比提高了一个数量级，由于在成像中使用了精细化的 fine 通道，单个时间片的最终数据量增加到 100GB 以上，所以在数据处理过程中，产生的数据可能会大大多于原始数据，特别是有特殊研究目的或有偶然的发现，可能会对原始数据进行二次甚至多次的深度分析。SKA 最初的观测将聚焦在几个核心科学项目（Key Science Project，简称 KSP）上面，比如脉冲星与宇宙再电离，而对这些数据的处理，均需要研发新的方法、算法、软件等，这些工作均需要在配套的平台上进行部署和测试，这也对 SRC 提出了很高的要求。

2.1　SRC 的网络需求

SRC 接收来自 SKA 天文台站的科学数据产品，并且为用户提供对科学数据产品的访问，以及配套的工具和处理方法 [5]。SRC 可以帮助科学家快

速地访问数据、处理数据以及发布数据。SRC 将支持通用的协作基础设施和中间件工具，以提供对数据、计算资源和存储资源的统一访问，而不受位置的影响。

SRC 将有助于充分利用其提供的计算资源和存储资源，因为随着 SKA 的持续运行，这些资源会保持持续的更新和升级，比如通过 CERN [8] 项目，可知即使 SRC 并不位于其设施的运行地理区域，项目也可以正常的运行。

SRC 除了主要提供从 SKA 站点移动数据、进行相关计算和存储外，还提供数据的后续处理，并允许相关的科研人员共享这些数据，如果政策和协议允许，这些数据也可以提供给其他科研人员共享和处理。

在网络带宽允许的情况下，SRC 将作为主要站点生成数据产品的异地备份设施，这样 SDP 站点就不需要保留该数据，但需要确保 SKA 对这些数据的访问权限。

综上，SRC 将提供的功能包括但不限于：

- 数据处理功能
- 数据存储功能
- 网络优化功能
- 异地备份功能

根据 SKAO 的规划，预计到 2025 年左右，只有少量 SKA 天线建成，每个台站的数据速率将不超过 20Gbps，随着建设的推进和科学观测的全面开展，2028 年左右 SKA 的数据流量将剧增，到 2030 年左右向 SRC 的输出速率将达到 100Gbps [19]。

3 数据传输软件与协议

传统模式下，观测数据会在本地进行存储和处理，但是随着 SKA 观测数据量的巨量增加，本地已经不足以支撑如此巨量的数据处理和分析，此时就需要通过高速网络传输到各个 SRC 进行分布式计算和存储。而影响长距离跨洲际的海量数据的传输主要包括：硬件设备、交换机设备、软件程序、使用的协议及对应的带宽等因素。

3.1 数据传输协议

在数据传输的测试中，使用了传统的提供端到端服务器的传输控制协议（Transmission Control Protocol，简称 TCP），用户数据报协议（User Datagram Protocol，简称 UDP），其中 TCP 稳定可靠且面向连接，UDP 简单易用不面向连接。由于 TCP 加入了各种安全保障功能，在实际执行的过程中会占用大量的系统开销，使得数据传输速度受到较大影响。随着数据流的增加，TCP 和 UDP 的性能均受到比较大的影响。

此时就出现了优化 TCP 协议的系列工作，大多从 TCP 的参数、窗口设置以及并发性来考虑和优化这些协议 [20]。最初的有针对单端优化 Tuning knobs，但都需要专业的管理员配置，且如果数据的双端不同时优化，可能会因为不同的网络情况而导致数据传输较慢。其他优化方法主要包括：修改 TCP 参数的 HighSpeed [21]，BiC [22]，FAST [23] 和 H-TCP [24]；并行化 TCP 协议的 PSockets [25] 和 GridFTP [26]；基于效率可靠 UDP 的 RBUDP [27]，Tsunami [28] 和 FOBS [29]；还有 SABUL [30]（UDT 协议的最早原型）及基于 UDP 的数据传输协议 [31]（UDP-based Data Transfer Protocol，简称 UDT）[1)]协议。

对于数据传输协议而言，需要满足流的特性、可靠性、支持单路和多路广播的功能；对于拥塞控制而言，也需要满足高效、公平及分布特性。

目前的这些协议也存在很多问题，比如对于改进的 TCP 就需要从内核层面进行编译部署，提高了测试门槛，参数的调优也需要人工干预。

对于 UDT 协议而言，该协议提供了应用层面的调用，提供与 TCP 类似的面向连接的功能，全新的协议设计和应用，新的拥塞控制算法，可配置的拥塞控制框架。第3.2.4节及第3.2.5节均使用了该协议。

安全远程登陆和文件传输协议（Secure Remote Login and File Transfer，简称 SSH[2)]）用于以加密的方式远程登陆其他计算机，该协议提供了几种不

1) www.udt.org

2) https://www.ssh.com/

同的加密认证选项，保证通讯的安全性，SSH 依赖于 openssh, 但代价是降低了整体的传输速率。

同时还有应用层的超文本传输协议（Hyper Text Transfer Protocol，简称 HTTP），HTTP 最初用于 Web 浏览器和服务器端的通信，现在也用于数据的传输工作。

3.2 数据传输软件

通过第3.1节的讨论，一般 TCP 协议用于专用线缆或短距离的数据传输；对于广域网或者长距离大多数使用 UDP 或其他协议。接下来介绍的基本为基于 UDP 的各类软件以及更高层的协议来实现的。

3.2.1 wget

wget[3)]是一个免费的命令行工具，基本在每个 Linux 发行版都默认安装。它使用 Internet 协议 HTTP、HTTPS、FTP 和 FTPS 来检索下载文件，可以通过前台或后台的方式工作。

wget 在低速网络上工作的非常好，可以保持获取数据的完整性，并且可以通过增量的方式重新更新获取文件。因为 HTTP 和 FTP 均带有时间戳，所以 wget 可以通过检索该属性来确认是否需要更新获取的文件。

wget 主要用于将数据通过 Internet 协议开放的用例，这个命令不需要对数据存储服务器或节点有访问权限就可以对公开数据进行下载。但这也会导致由于无法对网络进行优化而引起的性能问题。在同等网络条件下，相对于3.2.6数据传输的速率仅有 100Mbps。这种情况可以通过多线程的 wget 和 axel[4)]等来进行改进，在使用这种方法对南非 MeerKAT 的数据进行传输时，实测下来，多线程有相当大的速率提升，单线程只有 100Mbps，在开启 4 个线程的时候能够达到 500Mbps 的速率。但这里也取决于服务端对网络接口的限制。

3) http://www.gnu.org/software/wget/

4) https://www.axel.org/

5) https://iperf.fr

6) https://www.es.net/

7) https://www.lbl.gov/

3.2.2 scp

安全文件拷贝 (Secure File Copy，简称 scp) 用于在不同的计算机之间传输数据，该程序使用 SSH 协议，在大多数的 Linux 和 Unix 发行版默认安装。在测试中，*scp* 性能相对而言比较差，其中的加密校验增加了部分负荷，主要原因还是 openssh 的架构设计，其内部的 windowing 机制阻碍了快速连接的实现。特别是在长距离高延迟的情况下，这种影响尤为突出。SSH 通常将数据传输的窗口设定为 64KB，这在大数据流传输的情况下，是极大的瓶颈所在。通过动态定义窗口的缓存大小，可以将性能提升一个数量级以上 [32]，这种方法通过定义一个输入参数，来动态的调整窗口的大小，可以利用内核 TCP 栈的最佳效率 [20]。目前该方法的缺点在于，需要从源码编译，并且要保证数据发送端和接收端保持同样的版本。这对于网格化的多点传输而言，有比较大的困难，因为这需要具备管理员权限。在使用这种方法对澳大利亚 SKA 区域中心 Pawsey 超算中心数据服务器的数据进行传输时，单线程只有 100Mbps 的传输速率，这种情况下只能通过同时开启多个终端来加速数据的传输。

3.2.3 iperf3

IPerf3[5)]是一个开源的、支持多平台的工具，由美国能源网[6)]和劳伦斯伯克利国家实验室[7)]开发，用于主动测量 IP 网络上可实现的最大网络带宽。支持调整与时序、协议和缓冲区的各种相关参数，每次的测试都会显示测试的吞吐量、比特率等参数。IPerf3 支持目前常规的协议，比如 TCP、UDP 和 SCTP。

在本文研究中，主要使用 IPerf3 进行最初的相关测试，在 4.3 与西班牙 SKA 区域中心的评测中，也主要使用这款软件，以确保所有的参数均为最优状态。在网络传输中，传输窗口的大小对于测试非常重要，窗口指明了在接收应答信号 (acknowledge-

ment character，简称 ACK）之前允许传输的字节数目和数据包大小。窗口过小可能会导致直接连接的服务器和客户端频繁地切分网络数据包，导致传输速率被极大地影响。特别是随着 RTT 的增加，数据包和 ACK 之间的时间变长，如果窗口太小而无法容纳数据包，则发送方将被缩减以满足接收端的需求。

IPerf3 的用法为在其中一端使用 $iperf3-s$ 开启服务器程序，另外一端作为客户端进行测试，可以通过 IPerf3 提供的众多选项设定各个参数，比如

$$iperf3-c192.168.1.123-P4-t30-b10G-u$$

代表的是开始一个 UDP 传输，目标端地址为 192.168.1.123，并发开启 4 个线程，持续时间为 30 秒，带宽为 10Gbps。通过不同线程、及带宽的测试，可以动态调整所需要指定的窗口大小，以此完成对 UDP 协议的抖动、丢包率分析，TCP 和 UDP 协议的带宽适应性测试，从而找到最佳的传输性能参数。

3.2.4 NGAS

下一代归档系统（Next Generation Archive System，简称 NGAS）最初为欧洲南方天文台（European Southern Observatory，简称 ESO）的数据归档和分发而设计[33]，主要目标为整套系统可扩展并支持大数据的传输和检索。NGAS 的主要特点包括但不限于：

- 支持 Linux 不同的发行版操作系统
- 服务端使用 Python 开发
- 具体实现使用多线程 HTTP
- 基于 URL 的命令行接口
- 支持不同数据类型的插件架构
- 基于 XML 配置文件和消息传递

该技术后续也被用于 ALMA[34]、NRAO、MIT 等天文观测设备和观测台。目前 MWA 也使用了该技术[35]，并对其进行了定制，优化了 NGAS 的高吞吐数据输入、数据流管理、多层次数据存储和数据处理的缓存，以满足 MWA 数据 400MB/s 持续可见度数据的需求，该技术也用于 MWA 的数据分发工作。MWA 首先通过 NGAS 在台站以 400MB/s 的数据进行在线归档（Online Archive，简称 OA），后面通过 10Gbps 的网络传输到位于 Perth 超算的长期存储设备（Long-term Archive，简称 LTA），并同步发送到位于澳大利亚珀斯、美国 MIT 和新西兰 VUW 的数据处理中心。

NGAS 目前支持三种协议，分别为 HTTP、FTP 和 GridFT。基于 NGAS 和无限滑动窗口（Unlimited Sliding-Window，简称 USW）技术，通过 ZeroMQ 中间件技术，增强的 NGAS（Enhanced NGAS，简称 ENGAS）可以把 NGAS 的性能提升 3 到 12 倍。同时 ENGAS 还减少了通讯时间，改进了带宽的利用率，解决了远程同步的问题，从而达到更好的数据传输性能[36]。

3.2.5 jive5ab

jive5ab[37] 由欧洲 VLBI 联合所（Joint Institute for VLBI ERIC，简称 JIVE）为欧洲甚长基线干涉网络（European VLBI Network，简称 EVN）开发的一套软件。该软件旨在替换 Dimino，主要功能包括数据记录、读取和传输等。JIVE 为 EVN 的核心组织，主要工作为运行和开发 EVN 的观测数据及处理，EVN 的数据分散于全球 22 个望远镜和台站，这些望远镜最远距离 JIVE 有 10000 多公里。由于观测数据需要被统一传输到 JIVE 进行相关处理，所以需要一套实时数据传输的软件来进行运维，jive5ab 被用于将这些数据传输到相关处理中心进行处理，其中的工具 m5copy 基于 UDT 协议，可以开展长距离高速的数据传输及监测。目前上海天文台的天马 65 米射电望远镜与佘山 25 米射电望远镜作为 EVN 的两个参与台站，参与常规的 EVN 观测，观测的数据传输目前使用 jive5ab 来进行，EVN 的大多数台站（比如 Effelsberg、Yebes、Onsala 等）在 2017 年将 eVLBI 的带宽从 1Gbps 升级到了 2Gbps，天马 65 米射电望远镜与佘山 25 米射电望远镜也在 2018 年 5 月举行技术运行组会（Technical Operations Group，简称 TOG）之前，升级配置了 FlexBuff 用于通过 fila10g 支持 2 Gbps 的速率，后续随着中国科技网（China Sci-

ence and Technology Net，简称 CSTNet）的带宽升级，两个台站的数据传输带宽也从 1Gbps 升级到了 2Gbps，欧洲的大部分台站现在已经升级到了 4Gbps 带宽 [38]。

3.2.6 MWA download

全天虚拟天文台（The All-Sky Virtual Observatory，简称 ASVO）通过将澳大利亚所有的天文设施联合的方式，使研究人员和科学家可以访问所有的数据集。其中 ASVO-MWA 节点主要为 MWA 观测服务，通过在线服务和相应的工具为研究人员提供 MWA 的数据处理服务和获取，数据可以是原始可见度数据或校准可见度数据，可以通过后台搭载的软件来进行转换，同时 ASVO 节点强制实施元数据的完整性以及数据的访问权限设置。

当前 MWA 的数据下载主要通过 manta-ray-client[8)]来进行批量下载，使用基于 TCP 协议的应用层传输协议-超文本传输协议（HyperText Transfer Protocol，简称 HTTP），并通过多线程加速，默认为开启 4 个线程，数据块为 64KB。

ASVO-MWA 关于数据的原则称为 FAIR [39]，规则如下：

- 方便寻找的（Findable）：所有的 MWA 观测都有唯一的标识（使用 obsid 来查找）MWA ASVO 也提供与国际虚拟天文台联盟（International Virtual Observation Alliance，简称 IVOA）的表访问协议（Table Access Protocol，简称 TAP），方便与 IVOA 兼容的软件来访问观测信息
- 可访问的（Accessible），可以通过 TAP 协议便捷地访问
- 可互换的（Interoperable），MWA 可以生成两种标准的射电天文数据格式，即 CASA 测量集数据 [40]（Measurement Set，简称 MS）或者 fits 格式的数据 [41]
- 可重用的（Reusable），提供的数据包括元信息，比如观测、望远镜设置以及数据处理的细节

这些规则也适用于后续 SRC 对数据的需求。

8) https://github.com/MWATelescope/manta-ray-client

3.3 网络参数

通过带宽延迟（Bandwidth Delay Product，简称 BDP）的公式(1)可知，BDP 的大小决定了终端主机需要的缓冲量，很大程度上会影响数据传输速度，而最大传输速率很大程度上取决于网络数据包的窗口大小 [42]。

$$Bandwidth\ Delay = Bandwidth \times RTT \tag{1}$$

按照较优的 RTT，比如 40ms 为例，如果 Bandwidth 为 1Gbps，那么最佳的 BDP 应该设置为 4.77MB，同样如果 Bandwidth 为 10Gbps，最佳的 BDP 应该设置为 47.68MB。如果 RTT 为 10ms，在 1Gbps 带宽的时候，BDP 应该设置为此时的数据流为 1.19MB。

再结合 socket 缓冲区大小的计算(2)如下所示：

$$socket_buffer_size = 2 \times bandwidth \times delay \tag{2}$$

这个值可以通过 BSD setsocket 的选项 SO_SNDBUF 和 SO_RCVBUF 分别设定发送缓冲区和接收缓冲区的大小，这两个参数还会影响接收的窗口（Receiver Window，简称 RWND）。

系统初始的网络参数如下以及根据实际的输出传输测试，初始以及经过优化后的参数值如表1所示，其中各个配置位于 net 配置文件。

其中 net.core.netdev_max_backlog 用于控制 net_rx_action 软中断的最大值。

这些值系统默认情况下设置的比较小，需要根据内核的版本、数据的带宽以及可容忍的延迟来适当调整。

3.4 其他优化

对网络进行的调优也可以通过下面的操作来进行。使用大的数据块将改善性能，一般情况下推荐 8KB 的数据块。64KB 是更好的选择，这里需要网卡的硬件支持。另外创建并使用多个 socket 数据流

表 1 系统初始网络参数设置
Table 1 Network Parameter Setting

参数	初始值	优化值
core.rmem_max	212992	16777216
core.rmem_default	212992	16777216
core.wmem_max	212992	16777216
core.wmem_default	212992	16777216
ipv4.tcp_rmem	4096/87380/6291456	4096/87380/16777216
ipv4.tcp_wmem	4096/16382/4194304	4096/87380/16777216
core.netdev_max_backlog	1000	250000

可以极大地改善数据的传输性能。可以使用 iperf 的 -P 选项来进行实测。另外 bbcp 和 gridFTP 也支持并行数据流的传输。不过需要特别注意磁盘的性能，其性能也可能是数据读写的瓶颈。

如图所示，目前CSRC-P 的数据传输节点读写性能均在 20Gbps，满足高速传输的要求。

4 实验测试

本章节列举了CSRC-P与运行中的位于澳大利亚的 SKA 探路者项目默奇森宽视场阵 [17]、位于荷兰的 JIVE 研究以及与西班牙 SRC 进行的数据传输测试。传输测试详细阐述了使用的软件、方法及采用的相关优化等工作。

4.1 中澳实验

在CSRC-P 与澳大利亚的测试中，传输的数据主要来自 SKA 的先导阵列 MWA、ASKAP 以及用于脉冲星观测的 Parkes 望远镜。主要使用了三种传输方案，分别为 NGAS、ASVO 以及 scp。其中 NGAS 方案为最开始在CSRC-P与澳大利亚举办的大数据研讨会上[9)]开展，当时的网络由 CSTnet 和澳大利亚学术与研究网（Australian Academic Research Network，简称 AARNET）提供，带宽为 1Gbps。后续根据需要开通了常规的 2Gbps 带宽的网络以及突发的 5Gbps 的网络。

传输的数据格式主要包含 Measurement Set 数据，FITS 数据以及处理的结果文件。CASA MS 格式的数据有一个极大的数据文件和很多小文件，这种格式不利于数据的长距离传输，最好的办法是将其打包压缩为一个文件。而 FITS 文件主要为相关处理后的数据，数据量一般在每个文件 40GB ~ 60GB。处理的结果文件主要为一些文本文件和图像文件。

一种办法就是在 Pawsey 进行处理，后续直接在 Pawsey 的超算上进行下载，然后把数据处理的初步结果通过 Pawsey 超算的数据传输节点 hpc-data.pawsey.org.au 传输到CSRC-P，此时可以通过判断位于数据节点的 manifest 文件来判断文件是否更新，数据是否需要下载等信息，这种情况下因为使用的为 scp，所以每一组的数据速率仅仅能达到 150Mbps 左右。

CSRC-P最早利用 CSTNet 提供的 1G 网络使用 NGAS 开展了网络传输测试，当时的传输速率在 500Mbps~750Mbps，如图3,4所示：在 1Gbps 网络带宽的时候，使用 NGAS 进行数据的传输下载，在使用 1 个数据流传输的时候平均传输速率只有 514Mbps 的速率，通过增加 SOCKET 接收缓冲区的大小，并使用 4 个并行传输的线程，进行了 10 分钟约 64GB 的数据传输测试，平均传输速率达到了 738Mbps。考虑到从上海经由北京，然后通过公共互联网，到达澳大利亚珀斯，这个传输速率是比较理想的。

9) https://eridanus.net.au/?p=269

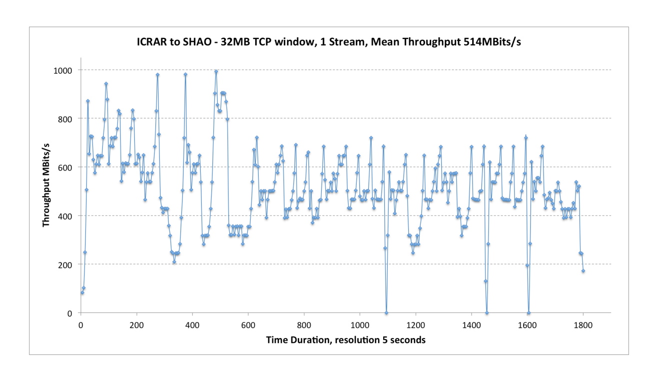


图 3 2017 年 5 月 19 日CSRC-P与 MWA 的数据流基准测试

Figure 3 Benchmark of NGAS using 1Gbps bandwidth on May 19 2017

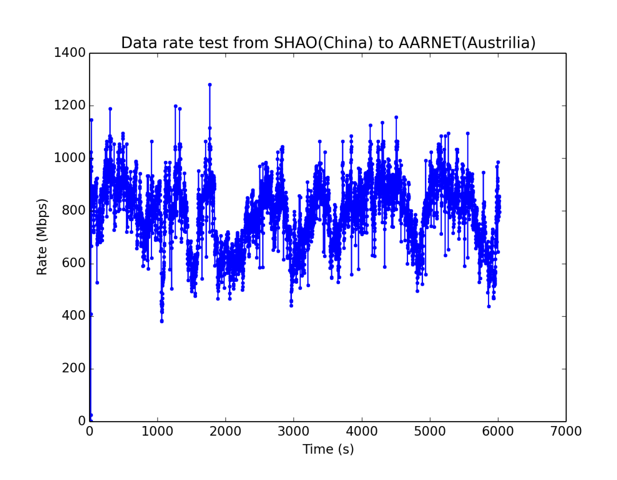


图 4 2017 年 8 月 28 日CSRC-P与 MWA 的 NGAS 数据流传输实验

Figure 4 Benchmark of NGAS between CSRC-P and ICRAR on August 28 2017

CSTNet 于 2020 年开通了 5G 带宽的测试条件，CSRC-P与 MWA 重新开展了数据传输，并使用全新的全天虚拟天文台（All-Sky Virtual Observatory，简称 ASVO）接口，通过 manta-ray-client 来进行数据下载，如图5数据的传输速率最高可达 4.5Gbps。目前 ASVO 的工作方式为，提交 job 到 ASVO 服务器，然后 ASVO 会将原始数据从长期存储的磁带中读回到硬盘缓冲区，在数据完整地读取到硬盘缓冲区以后，才可以进行下载，所以在提交新的数据下载请求时，可以看到会有一段时间的等待，此时即为数据的准备时间。而从 2021 年 10 月 1 日 02 时开始到 09 时左右的数据速率为 0，后续分析得知，该段时间 ASVO 节点处于维护阶段，暂停所有的数据处理，后续的数据传输恢复正常。

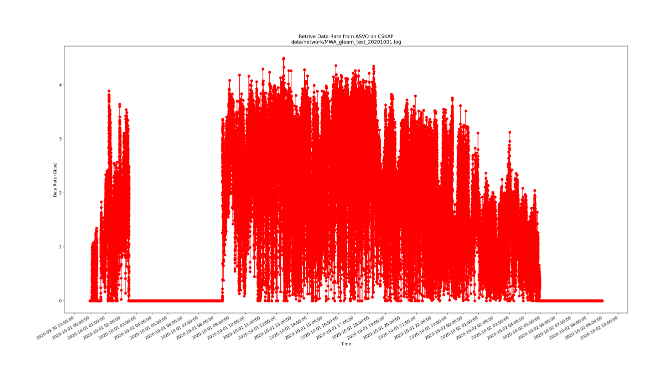


图 5 2020 年 10 月 1 日CSRC-PChina-ASVO 测试

Figure 5 Benchmark of ASVO using 5Gbps bandwidth on October 1 2020

4.2 中荷实验

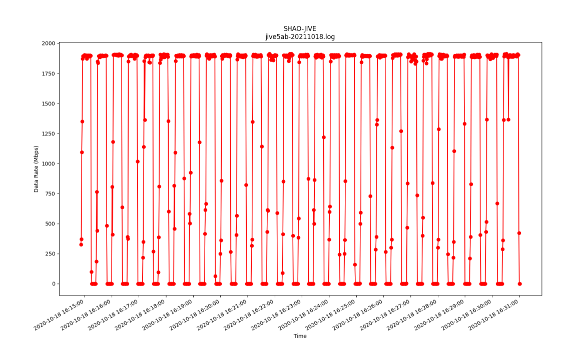


图 6 2021 年 10 月 18 日 China-Jive 数据常规传输 2Gbps 带宽

Figure 6 Data transfer between China and JIVE on October 18 2021 using 2Gbps bandwidth

上海天文台有 2 台射电望远镜参加 EVN 观测，会有常规的数据传输工作。在与 JIVE 进行的数据传输中，传输的数据主要为 EVN 的观测数据，数据格式当前主要为 Mark5B，根据观测计划的不同，单个数据文件的大小在 10GB ~150GB，基于 UDT 协议使用 jive5ab 来传输。当前的网络带宽为 2Gbps，如图6所示，数据传输的平均速率在 1.8Gbps 左右，

带宽利用率在 90% 左右。数据传输已经经过了优化，本次的观测波动主要在首次读取文件的时刻。

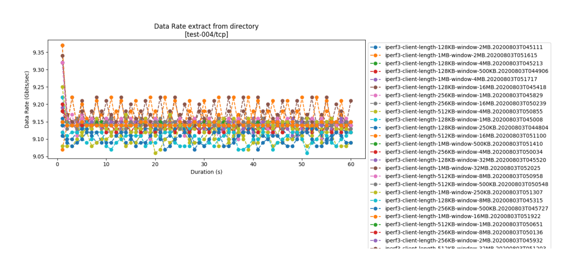


图 8　2020 年 8 月 6 日 10Gbps 不同 TCP 窗口大小测试
Figure 8　Benchmark of different TCP length using 10Gbps bandwidth on August 6 2020

4.3　中西实验

2020 年 7 月开始，CSRC-P 与西班牙安达卢西亚天体物理研究所（The Instituto Astrofísica Andalucía，简称 IAA）开展了关于 SRC 之间的网络测试研究，用于分析出在跨洲际的情况下，影响网络性能的因素。主要分析了短距离 10Gbps 及 25Gbps 带宽情况下，基于 UDP 和 TCP 协议影响数据传输的情况。由图7,8, 9,10可知，在 10Gbps 带宽的情况下，TCP 数据包在 256KB 以上，其传输速率基本保持比较稳定的状态，TCP 窗口对速率的影响在 2% 左右；在 25Gbps 带宽的情况下，TCP 数据包在 256KB 及 512KB 时，其传输速率可以达到较好的状态，TCP 窗口在 4MB，数据包为 512KB 时，速率最高，基本可以到到 19Gbps，对速率的影响较大，可以达到在 10% 左右。在对CSRC-P与 IAA 进行数据传输时，在网络带宽为 1Gbps 的情况下传输的测试速率仅仅在 150Mbps 左右。分析可能的情况为受到了中间路由的限制。

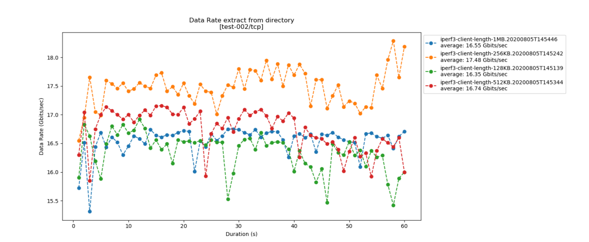


图 9　2020 年 8 月 6 日 25Gbps 不同 TCP 长度测试
Figure 9　Benchmark of different TCP length using 25Gbps bandwidth on August 6 2020

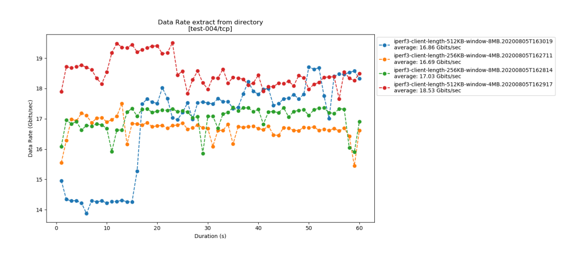


图 10　2020 年 8 月 6 日 25Gbps 不同 TCP 窗口大小测试
Figure 10　Benchmark of different TCP length using 25Gbps bandwidth on August 6 2020

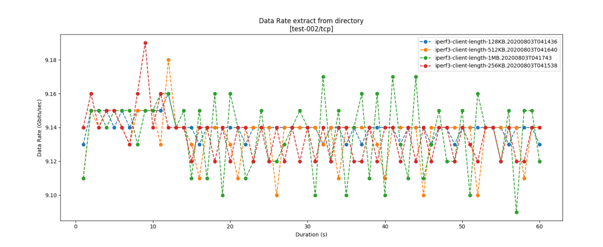


图 7　2020 年 8 月 6 日 10Gbps 不同 TCP 长度测试
Figure 7　Benchmark of different TCP length using 10Gbps bandwidth on August 6 2020

5　讨论

相对于 UDP 而言，TCP 提供了更多的控制机制，比如 sliding 窗口，数据重传，Nagle 算法，拥塞控制等。对于短距离传输而言，UDP 是比较好的通讯协议，但是对长距离，比如洲际之间的传输，大多仍需使用 TCP 传输协议，或者也可以在 UDP 的基础上增加一些可靠性的机制。下一步将进一步对 TCP 协议的高速存储技术进行研究和测试，包括对多核多进程开展相关研究。当前 40Gbps/100Gbps

的网络也在快速研发中，由于其价格昂贵，基本用于洲际的骨干网和数据中心，相信随着新技术的发展，硬件也会逐步升级。当前北京与上海已经具备了 100Gbps 的高速网络，后续将尽快开展相关的优化工作，提升 100Gbps 网络的性能，以应对未来的挑战。

6 总结与展望

SKA 的科学潜力是前所未有的，也是国际天文学界的最高优先事项之一。同样 SKA 产生数据的规模、速率以及复杂度，都对当前业界的数据管理、网络和计算提出了很大的挑战。通过对数据速率、存储和处理等流程的估算，每个 SKA 望远镜台站每年有超过 300PB 的数据需要传输到各个区域中心，观测台站到每个 SRC 需要具备 100Gbps 的网络，这些工作十分依赖于国家的骨干网络设施、设备和相应的软件。随着技术的进步和 SKA 项目的稳步推进，新的技术和方法也将对业界产生比较好的推动作用。

致谢 向评审人和对该文有帮助的人士表示谢意. 本研究使用了中国 SKA 区域中心原型机的资源。

参考文献

1 武向平. 中国 SKA 科学白皮书. 北京：科学出版社，2017

2 An T, Wu X P, Hong X Y, et al. Science Applications and Challenges of SKA Big Data(in Chinese). Bulletin of the Chinese Academy of Sciences. 2018，8: 871–876[安涛, 武向平, 洪晓瑜, 等. SKA 大数据的科学应用和挑战. 中国科学院院刊，2018，8: 871–876]

3 Guo S G, Zheng X Y, Mao Y F, et al. Scheme and Prospect of the SKA Big Data Transferring(in Chinese). E-science Technology & Application, 2018，9(3): 3–13 [郭绍光，郑小盈，毛羽丰，等. SKA 海量数据传输的方案及展望. 科研信息化技术与应用，2018，9(3): 3–13]

4 Quinn P, van Haarlem M, An T, et al. SKA Regional Centres White Paper v1.0. Technical Report, Square Kilometre Array Organisation. 2020

5 Chrysostomou A, the SRCCG. SKA REGIONAL CENTRES:BACKGROUND AND FRAMEWORK. Techinical Report, Square Kilometre Array Organisation. 2017

6 An T, Wu X C, Lao B Q, et al. Status and progress of China SKA Regional Centre prototype. arxiv:2206.13022

7 An T, Wu X P, Hong X Y, SKA data take centre stage in China. Nat Astron, 2019, 3: 1030–1030

8 Shu T, Todorovic S, Zhu S. CERN: confidence-energy recurrent network for group activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. eprint arXiv:1704.03058

9 Guo S G, Lu Y, An T, et al. Scientific data flow and array simulation analysis for the SKA-1 era(in Chinese). in prepare (2022) [郭绍光，陆扬，安涛，等. 面向 SKA-1 时代的科学数据流及阵列模拟分析. 准备中 (2022).

10 Lao B Q, Zhang Y K, An T, et al. Software Platform on China SKA Regional Center Prototype System(in Chinese).ChinaXiv:202206.00173. [劳保强，张迎康，安涛，等.(2022). 中国 SKA 区域中心原型系统 – 软件平台.ChinaXiv:202206.00173]

11 Xu Z J, An T, Guo S G, et al. A machine learning dataset for FRB detection in raw data(in Chinese).ChinaXiv:T202206.00321.[徐志骏， 安涛， 郭绍光， 等.(2022). 一个面向原始数据搜寻的快速射电暴数据集.ChinaXiv:T202206.00321]

12 Wei J W, Zhang C F, Zhang Z L, et al. Parallel optimization of the pulsar search pipeline on China SKA Regional Centre Prototype (in Chinese). ChinaXiv:T202206.00297. [韦建文，张晨飞，张仲莉，等.(2022). 低频射电脉冲星搜索的性能优化方法. ChinaXiv:T202206.00297]

13 Wei J W, Zhang C F, Lao B Q, et al. Optimization of parallel processing of Square Kilometre Array low frequency imaging pipeline (in Chinese). ChinaXiv:T202206.00292. [韦建文，张晨飞，劳保强，等.(2022).SKA 低频成像管线并行优化化.ChinaXiv:T202206.00292]

14 Bolton R C, the SRCCG. SKA REGIONAL CENTRE REQUIREMENTS. Technical Report, Square Kilometre Array Organisation. 2019
15 Ivashina M V, Ardenne A V, Bregman J D, et al. Activities for the square kilometer array (SKA) in Europe. In: 4th International Conference on Antenna Theory and Techniques (Cat. No. 03EX699). IEEE, 2003. Vol. 2
16 Santander-Vela, Juande, Lorenzo Pivetta, and Nick Rees. "Status of the Square Kilometre Array." Proc. of International Conference on Accelerator and Large Experimental Control Systems (ICALEPCS' 17), Barcelona, Spain, 8-13 October 2017. 2017.
17 Whitney A, Booler T, Bowman J, et al. The Murchison Widefield Array (MWA): Current Status and Plans. In: American Astronomical Society, AAS Meeting #218. Bulletin of the American Astronomical Society, 2011. Vol. 43, id.132.07
18 Wayth R, Tingay S, Trott C. The phase II Murchison widefield array: design overview. Publications of the Astronomical Society of Australia,2018,35:id.e033
19 Ford, D, Bolton, R, Colegate, T, et al. (2010). The SKA costing and design tool. SKA Memo Series, Memo, 120, 1-31.
20 Semke, Jeffrey, Jamshid Mahdavi, and Matthew Mathis. "Automatic TCP buffer tuning." Proceedings of the ACM SIGCOMM'98 conference on Applications, technologies, architectures, and protocols for computer communication. 1998.
21 Imran A, Zulkar N, Md S Q, et al. OneDataShare:A Vision for Cloud-hosted Data Transfer Scheduling and Optimization as a Service. 2019. eprint arXiv:1712.02944
22 Xu, Lisong, Khaled Harfoush, and Injong Rhee. "Binary increase congestion control (BIC) for fast long-distance networks." IEEE INFOCOM 2004. Vol. 4. IEEE, 2004.
23 Jin, Cheng, David X. Wei, and Steven H. Low. "FAST TCP: motivation, architecture, algorithms, performance." IEEE INFOCOM 2004. Vol. 4. IEEE, 2004.
24 Leith, Doug, and Robert Shorten. "H-TCP: TCP congestion control for high bandwidth-delay product paths." draft-leith-tcp-htcp-06 (work in progress) (2008).
25 Sivakumar, Harimath, Stuart Bailey, and Robert L. Grossman. "PSockets: The case for application-level network striping for data intensive applications using high speed wide area networks." SC'00: Proceedings of the 2000 ACM/IEEE Conference on Supercomputing. IEEE, 2000.
26 Allcock, William, et al. "The Globus striped GridFTP framework and server." SC'05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing. IEEE, 2005.
27 He, Eric, et al. "Reliable blast UDP: Predictable high performance bulk data transfer." Proceedings. IEEE International Conference on Cluster Computing. IEEE, 2002.
28 Meiss, Mark R. "Tsunami: A high-speed rate-controlled protocol for file transfer." Indiana University (2004).
29 Dickens, Phillip M. "FOBS: A lightweight communication protocol for grid computing." European Conference on Parallel Processing. Springer, Berlin, Heidelberg, 2003.
30 Gu, Yunhong, and Robert Grossman. "SABUL: A transport protocol for grid computing." Journal of grid Computing 1.4 (2003): 377-386.
31 Gu, Yunhong, and Robert L. Grossman. "UDT: UDP-based data transfer for high-speed wide area networks." Computer Networks 51.7 (2007): 1777-1799.
32 Rapier, Chris, and M. Stevens. "Application Layer Network Window Management in the SSH Protocol." Supercomputing Conference. 2004.
33 Wicenec A, Knudstrup J, Johnston S. ESO's Next Generation Archive System. The Messenger, 2011, 106: 11–13
34 Wicenec A, Farrow S, Gaudet S, et al. The ALMA Archive: A Centralized System for Information Services. In: Astronomical Data Analysis Software and Systems (ADASS) XIII. San Francisco: Astronomical Society of the Pacific, 2004. 93–96
35 Wu C, Wicenec A, Pallot D, et al. Optimising NGAS for the MWA Archive. Experimental Astronomy, 2013, 36(3): 679–694.
36 Shi C, Deng H, Wang F, et al. Enhanced remote astronomical archive system based on the file-level Unlimited Sliding-Window technique. Research in Astronomy and Astrophysics, 2021, 21(10): 253–260
37 Harro Verkouter. jive5ab documentation Version 2.5, 2020
38 EUROPEAN VLBI NETWORK biennial report 2017–2018. Technical Report, EVN. 2018
39 O'Toole S, Tocknell J. FAIR standards for astronomical data. 2022. arXiv preprint arXiv:2203.10710
40 McMullin, J. P., et al. "ASP Conf. Ser. Vol. 376, Astronomical Data Analysis Software and Systems XVI." (2007): 127.
41 Greisen, E. W. Aips memo series, AIPS FITS File Format. AIPS Memo 117, 2019.

42 Stevens W. TCP/IP Illustrated. Volume 1: The Protocols. Addison-Wesley, 1994

Progress and Prospect of transcontinental high-speed data transmission at SKA Regional Center in China

GUO ShaoGuang[1*], AN Tao[1], XU ZhiJun[1], LAO BaoQiang[1], LU Yang[1], LU WeiJia[1] & WU Xiaocong[1]

1. Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Shanghai 200030, China

The Square Kilometer Array (SKA) is the largest radio telescope, and the data generated by its observations will be transmitted from Australia and South Africa to the scientific data processing center about one hundred kilometers away at first, and then distributed to various SKA Regional Centres(SRC) with a distance of tens of thousands of kilometers through high-speed network. In the SKA Phase One (SKA1) stage with a scale of 10% of SKA, it is estimated that about 750PB of data needs to be distributed to each SRC through a network of at least 100Gbps each year. Such high network bandwidth and data scale bring great challenges to data transmission and distribution. This paper analyzes different network protocols such as TCP/UDP/HTTP and uses different software in the field of radio astronomy for testing and research, and then the optimal transmission scheme parameters under the current infrastructure of 10Gbps network are obtained. In this paper, the factors affecting high-speed transmission are discussed, and the corresponding performance optimization strategies are given. Before the real observation data of SKA1 is generated, it will provide the technical foundation for the network construction and layout of China's SKA regional center. The technical details and methods described are available for reference and use in relevant scientific applications. Finally, the challenges of future SKA network requirements are discussed and prospected.

Square Kilometre Array , SKA Regional Centre, high speed network, Data transfer

PACS: 47.27.-i, 47.27.Eq, 47.27.Nz, 47.40.Ki, 47.85.Gj

doi: ??